AUTOMATED DETERMINATION OF SUBLANGUAGE SYNTACTIC USAGE

Ralph Grishman and Ngo Thanh Nhan

Courant Institute of Mathematical Sciences
New York University
New York, NY 10012

Elaine Marsh

Navy Center for Applied Research in Artificial Intelligence
Naval Research Laboratory
Washington, DC 20375

Lynette Hirschman

Research and Development Division
System Development Corporation / A Burroughs Company
Paoli, PA 19301

Abstract

Sublanguages differ from each other, and from the "standard language", in their syntactic, semantic, and discourse properties. Understanding these differences is important if we are to improve our ability to process these sublanguages. We have developed a semiautomatic procedure for identifying sublanguage syntactic usage from a sample of text in the sublanguage. We describe the results of applying this procedure to three text samples: two sets of medical documents and a set of equipment failure messages.

Introduction

A sublanguage is the form of natural language used by a community of specialists in discussing a restricted domain. Sublanguages differ from each other, and from the "standard language", in their syntactic, semantic, and discourse properties. We describe here some recent work on (semi-)automatically determining the syntactic properties of several sublanguages. This work is part of a larger effort aimed at improving the techniques for parsing sublanguages.

If we examine a variety of scientific and technical sublanguages, we will encounter most of the constructs of the standard language, plus a number of syntactic extensions. For example, "report" sublanguages, such as are used in medical summaries and equipment failure summaries, include both full sentences and a number of fragment forms [Marsh 1983]. Specific sublanguages differ in their usage of these syntactic constructs [Kittredge 1982, Lehrberger 1982].

Identifying these differences is important in understanding how sublanguages differ from the language as a whole. It also has immediate practical benefits, since it allows us to trim our grammar to fit the specific sublanguage we are processing. This can significantly speed up the analysis process and block some spurious parses which would be obtained with a grammar of overly broad coverage.

Determining Syntactic Usage

Unfortunately, acquiring the data about syntactic usage can be very tedious, inasmuch as it requires the analysis of hundreds (or even thousands) of sentences for each new sublanguage to be processed. We have therefore chosen to automate this process.

We are fortunate to have available to us a very broad coverage English grammar, the Linguistic String Grammar [Sager 1981], which has been extended to include the sentence fragments of certain medical and equipment failure reports [Marsh 1983]. The grammar consists of a context-free component augmented by procedural restrictions which capture various syntactic and sublanguage semantic constraints. The context-free component is stated in terms of grammatical categories such as noun, tensed verb, and adjective.

To begin the analysis process, a sample corpus is parsed using this grammar. The file of generated parses is reviewed manually to eliminate incorrect parses. The remaining parses are then fed to a program which counts -- for each parse tree and cumulatively for the entire file -- the number of times that each production in the context-free component of the grammar was applied in building the tree. This yields a "trimmed" context-free grammar for the sublanguage (consisting of those productions used one or more times), along with frequency information on the various productions.

This process was initially applied to text samples from two sublanguages. The first is a set of six patient documents (including patient history, examination, and plan of treatment). The second is a set of electrical equipment failure reports called "CASREPs", a class of operational report used by the U. S. Navy [Froscher 1983]. The parse file for the patient documents had correct parses for 236 sentences (and sentence fragments); the file for the CASREPs had correct parses for 123 sentences. We have recently applied the process to a third text sample, drawn from a sublanguage very similar to the first: a set of five hospital "discharge summaries", which include patient histories, examinations, and summaries of the course of treatment in the hospital. This last sample included correct parses for 310 sentences.

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **JUL 1994**	2. REPORT TYPE	3. DATES COVERED **00-00-1994 to 00-00-1994**

4. TITLE AND SUBTITLE **Automated Determination of Sublanguage Syntactic Usage**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Naval Research Laboratory,Navy Center for Applied Research in Atificial Intelligence,Washington,DC,20375**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S)
	11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

Proceedings of the 22nd Annual Meeting of the Association for Computational Linguistics, held 2-6 July, 1984 in Stanford, CA.

14. ABSTRACT

Sublanguages _differ from each other, and from the "standard Ian~age, in their syntactic, semantic, and discourse vrolx:rties. Understanding these differences is important’if -we are to improve our ability to process these sublanguages. We have developed a sen~.’- automatic ~ure for identifying sublangnage syntact/c usage from a sample of text in the sublanguage..We describe the results of applying this procedure to taree text samples: two sets of medical documents and a set of equipment failure me~ages.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**5**	

Results

The trimmed grammars produced from the three sublanguage text samples were of comparable size. The grammar produced from the first set of patient documents contained 129 non-terminal symbols and 248 productions; the grammar from the second set (the "discharge summaries") was slightly larger, with 134 non-terminals and 282 productions. The grammar for the CASREP sublanguage was slightly smaller, with 124 non-terminals and 220 productions (this is probably a reflection of the smaller size of the CASREP text sample). These figures compare with 255 non-terminal symbols and 744 productions in the "medical records" grammar used by the New York University Linguistic String Project (the "medical records" grammar is the Linguistic String Project English Grammar with extensions for sentence fragments and other, sublanguage specific, constructs, and with a few options deleted).

Figures 1 and 2 show the cumulative growth in the size of the trimmed grammars for the three sublanguages as a function of the number of sentences in the sample. In Figure 1 we plot the number of non-terminal symbols in the grammar as a function of sample size; in Figure 2, the number of productions in the grammar as a function of sample size. Note that the curves for the two medical sublanguages (curves A and B) have pretty much flattened out toward the end, indicating that, by that point, the trimmed grammar covers a very large fraction of the sentences in the sublanguage. (Some of the jumps in the growth curves for the medical grammars reflect the division of the patient documents into sections (history, physical exam, lab tests, etc.) with different syntactic characteristics. For the first few documents, when a new section begins, constructs are encountered which did not appear in prior sections, thus producing a jump in the curve.)

The sublanguage grammars are substantially smaller than the full English grammar, reflecting the more limited range of modifiers and complements in these sublanguages. While the full grammar has 67 options for sentence object, the sublanguage grammars have substantially restricted usages: each of the three sublanguage grammars has only 14 object options. Further, the grammars greatly overlap, so that the three grammars combined contain only 20 different object options. While sentential complements of nouns are available in the full grammar, there are no instances of such constructions in either medical sublanguage, and only one instance in the CASREP sublanguage. The range of modifiers is also much restricted in the sublanguage grammars as compared to the full grammar. 15 options for sentential modifiers are available in the full grammar. These are restricted to 9 in the first medical sample, 11 in the second, and 8 in the equipment failure sublanguage. Similarly, the full English grammar has 21 options for right modifiers of nouns; the sublanguage grammars had fewer, 11 in the first medical sample, 10 in the second, and 7 in the CASREP sublanguage. Here the sublanguage grammars overlap almost completely: only 12 different right modifiers of noun are represented in the three grammars combined.

Among the options occurring in all the sublanguage grammars, their relative frequency varies according to the domain of the text. For example, the frequency of prepositional phrases as right modifiers of nouns (measured as instances per sentence or sentence fragment) was 0.36 and 0.46 for the two medical samples, as compared to 0.77 for the CASREPs. More striking was the frequency of noun phrases with nouns as modifiers of other nouns: 0.20 and 0.32 for the two medical samples, versus 0.80 for the CASREPs.

We reparsed some of the sentences from the first set of medical documents with the trimmed grammar and, as expected, observed a considerable speed-up. The Linguistic String Parser uses a top-down parsing algorithm with backtracking. Accordingly, for short, simple sentences which require little backtracking there was only a small gain in processing speed (about 25%). For long, complex sentences, however, which require extensive backtracking, the speed-up (by roughly a factor of 3) was approximately proportional to the reduction in the number of productions. In addition, the frequency of bad parses decreased slightly (by <3%) with the trimmed grammar (because some of the bad parses involved syntactic constructs which did not appear in any correct parse in the sublanguage sample).

Discussion

As natural language interfaces become more mature, their portability -- the ability to move an interface to a new domain and sublanguage -- is becoming increasingly important. At a minimum, portability requires us to isolate the domain dependent information in a natural language system [Grosz 1983, Grishman 1983]. A more ambitious goal is to provide a *discovery procedure* for this information -- a procedure which can determine the domain dependent information from sample texts in the sublanguage. The techniques described above provide a partial, semi-automatic discovery procedure for the syntactic usages of a sublanguage.* By applying these techniques to a small sublanguage sample, we can adapt a broad-coverage grammar to the syntax of a particular sublanguage. Subsequent text from this sublanguage can then be processed more efficiently.

We are currently extending this work in two directions. For sentences with two or more parses which satisfy both the syntactic and the sublanguage selectional (semantic) constraints, we intend to try using the *frequency* information gathered for productions to select a parse. We shall determine whether there is a correlation in these cases between the correct parse and the parse involving the more frequent syntactic constructs.** Second, we are using a similar approach to develop a discovery procedure for sublanguage selectional patterns. We are collecting, from the same sublanguage samples, statistics on the frequency of co-occurrence of particular sublanguage (semantic) classes in subject-verb-object and host-adjunct relations, and are using this data as input to

* Partial, because it cannot identify new extensions to the base grammar; semi-automatic, because the parses produced with the broad-coverage grammar must be manually reviewed.

** Some small experiments of this type have been done with a Japanese grammar [Nagao 1982] with limited success. Because of the very different nature of the grammar, however, it is not clear whether this has any implications for our experiments.

the grammar's sublanguage selectional restrictions.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grants No. MCS-82-02373 and MCS-82-02397.

References

[Froscher 1983] Froscher, J.; Grishman, R.; Bachenko, J.; Marsh, E. "A linguistically motivated approach to automated analysis of military messages." To appear in *Proc. 1983 Conf. on Artificial Intelligence*, Rochester, MI, April, 1983.

[Grishman 1983] Grishman, R.; Hirschman, L.; Friedman, C. "Isolating domain dependencies in natural language interfaces." *Proc. Conf. Applied Natural Language Processing*, 46-53, Assn. for Computational Linguistics, 1983.

[Grosz 1983] Grosz, B. "TEAM: a transportable natural-language interface system." *Proc. Conf. Applied Natural Language Processing*, 39-45, Assn. for Computational Linguistics, 1983.

[Kittredge 1982] Kittredge, R. "Variation and homogeneity of sublanguages." In *Sublanguage: studies of language in restricted semantic domains*, ed. R. Kittredge and J. Lehrberger. Berlin & New York: Walter de Gruyter; 1982.

[Lehrberger 1982] Lehrberger, J. "Automatic translation and the concept of sublanguage." In *Sublanguage: studies of language in restricted semantic domains*, ed. R. Kittredge and J. Lehrberger. Berlin & New York: Walter de Gruyter; 1982.

[Marsh 1983] Marsh, E.. "Utilizing domain-specific information for processing compact text." *Proc. Conf. Applied Natural Language Processing*, 99-103, Assn. for Computational Linguistics, 1983.

[Nagao 1982] Nagao, M.; Nakamura, J. "A parser which learns the application order of rewriting rules." *Proc. COLING 82*, 253-258.

[Sager 1981] Sager, N. *Natural Language Information Processing*. Reading, MA: Addison-Wesley; 1981.

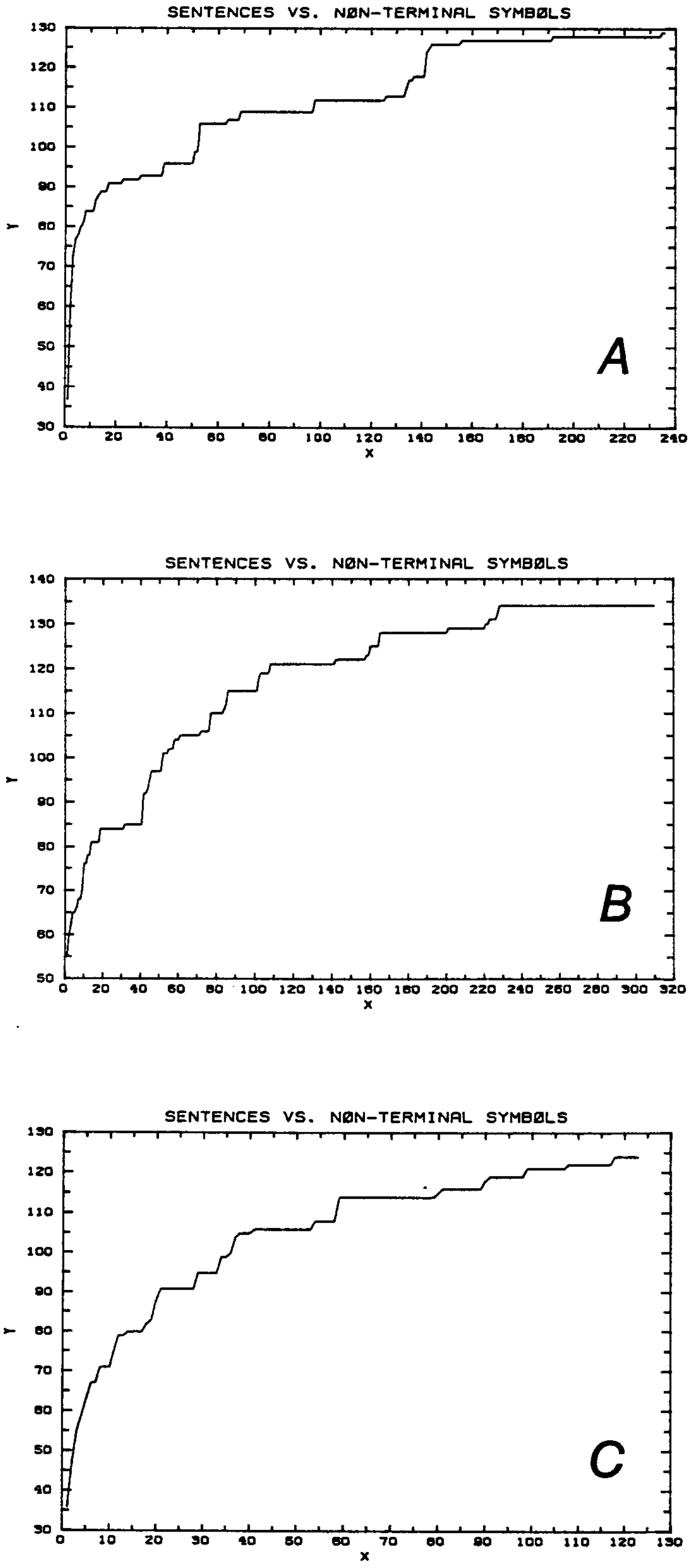

Figure 1. Growth in the size of the grammar as a function of the size of the text sample. X = the number of sentences (and sentence fragments) in the text sample; Y = the number of non-terminal symbols in the context-free component of the grammar.

Graph A: first set of patient documents

Graph B: second set of patient documents ("discharge summaries")

Graph C: equipment failure messages ("CASREPs")

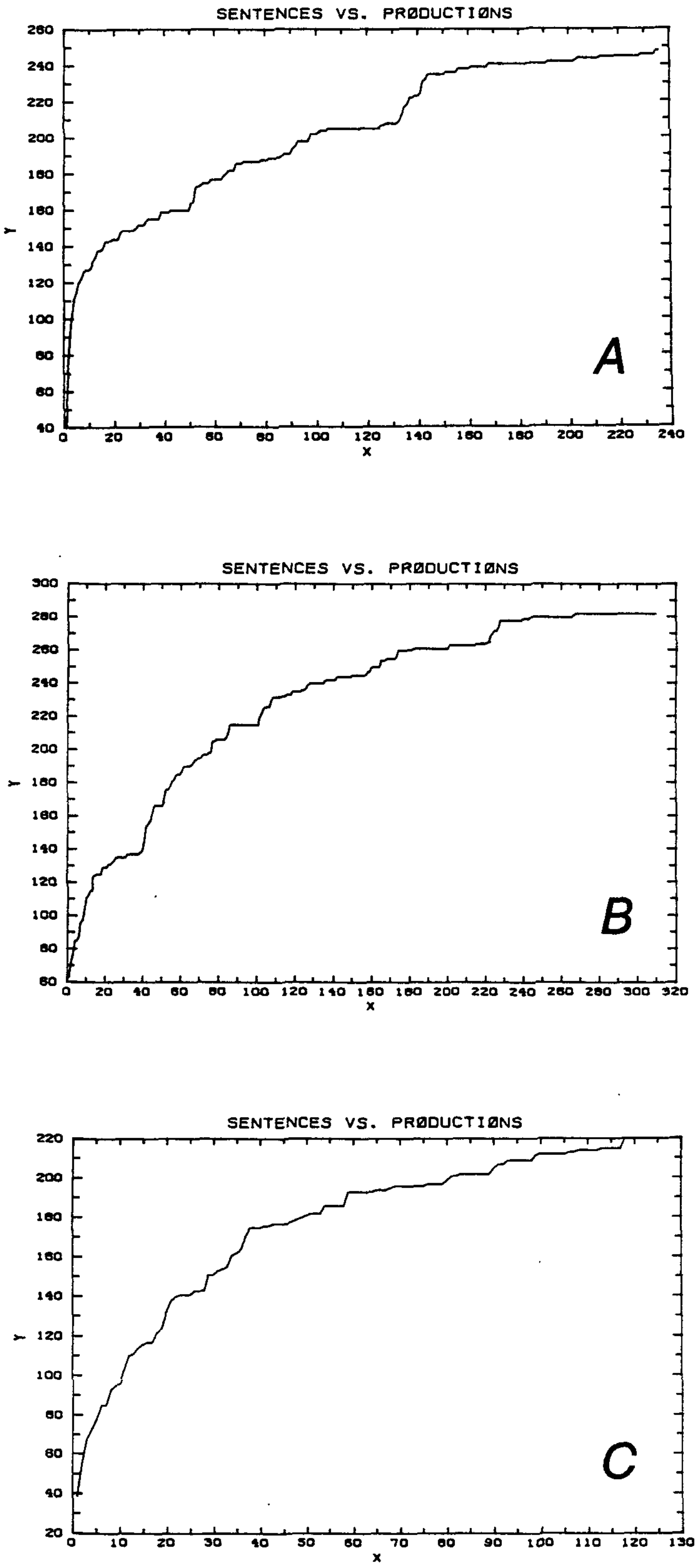

Figure 2. Growth in the size of the grammar as a function of the size of the text sample. X = the number of sentences (and sentence fragments) in the text sample; Y = the number of productions in the context-free component of the grammar.

Graph A: first set of patient documents

Graph B: second set of patient documents ("discharge summaries")

Graph C: equipment failure messages ("CASREPs")